

Weekly Report (2014.09.15~09.21)

Done

1) 小结下关于“将数据表格直接拍照，自动转换成数据，并自定义（visketch）可视化结果”的项目：

1. 愿景：用户通过手机拍照，可直接将带有数据表格的图像转换成数据（支持用户进行数据纠错），并在此基础上快速地完成可视设计（这个步骤可以首先根据数据表格类型，自动生成可视化方案，然后支持用户后续调整），帮助用户理解数据并便于分享。对于多数普通用户，整个过程可以全自动完成，无需进行交互。

2. 可行性：

图片->数据步骤：微软的 OneNote 产品里面集成了 Office lens 插件（

视频：http://v.youku.com/v_show/id_XNzg0NjlxNjg0.html

博客介绍：http://blog.sina.com.cn/s/blog_4caedc7a0102f0vt.html

），可以将打印的文件或者白板上的草图直接转换成 office 文档（word、ppt 等），这个工具目前还不确定是否有接口。如果没有，则需使用其他 OCR 库（比如：

<https://github.com/gali8/Tesseract-OCR-iOS>,

<https://dev.myscript.com/technology/text/>）对数据进行识别（由于可能没有图片矫正、增强，效果不如 Office lens），至于怎么把识别的结果进一步转换成表格，还要找找看有无现成方法。

数据->可视化步骤：可以集成小马他们之前做的 visketch 项目，应该无障碍。

具体的下周一起讨论

3. 参考文献：

<http://hci.stanford.edu/publications/2011/revision/revision-UIST2011.pdf>

Jeffrey Heer 等人对已有的可视化结果直接进行识别，转化成数据然后进行重新可视化，旨在对目前很多不合理的可视化结果进行自动优化。我们这个项目和他的区别在于：我们是直接对数据进行识别，且做到移动端（手机上），更具实用性、灵活性。至少在引用这篇文章的文章里面，我没有看到类似的工作。

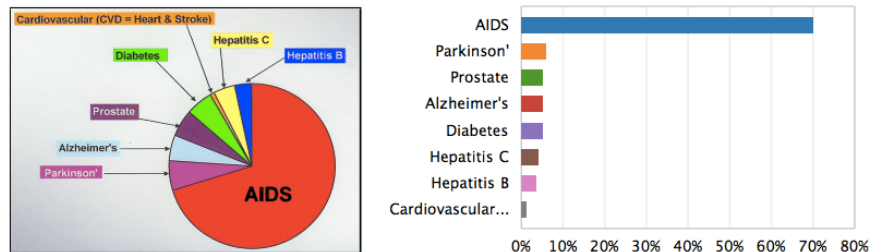


Figure 1: Chart Redesign. Left: A pie chart of NIH expenses per condition-related death. The chart suffers from random sorting, highly saturated colors, and erratic label placement. Right: Plotting the data as a sorted bar chart enables more accurate comparisons of data values [6, 20].

2) 除了 storyline 之外，看了两篇文章：

1. *Using Topological Analysis to Support Event-Guided Exploration in Urban Data*

这篇文章针对时空数据，在时间维度上，将数据切片，对每片数据用标量函数（原文 scalar function）去逼近，通过寻找标量函数上的极值点去提取事件。进而在相邻时间片中将相似的事件进行关联，从而提取出跨越多个时间片的事件。最后将事件投影

到 **density** (单位时间发生的事件数量) 和 **range** (事件持续时间) 定义的二维空间上, 同时定义了事件的重要性并进行排序, 帮助用户快速定位不同类型的事件。

2. Travel Time Estimation of a Path using Sparse Trajectories

这篇文章针对车辆的 GPS 轨迹数据, 可以预测经过任意路径所需的时间。本文首相将路网分割成一段一段 (**segments**), 将时间也分成一个个时间段 (**time slots**)。一辆车在一个时间段进过一个路段会形成一个三维的数据 (**t, s, v**) (分别对应 **time slot**, **segment**, **velocity**), 本文继而将这个数据看成一个三维张量, 并用张量分解的方法去对确实的数据进行补全。

对于一次查询, 即查询经过某条路径 **P** 需要的时间, 本文不是简单地对相应的 **segments** 作平均, 而是提出了一个能量优化方程, 通过最小化组成 **P** 的 **segments** 的协方差来优化 **P** 的 **segment** 组成方案。

对于卡口类型的数据 (测速点固定), 这种方法是不适用的: 首先它相对于卡口数据可能更为稠密 (平均每辆车每 96s 一个点); 其次, GPS 数据的分散性导致了几乎在每个 **segment** 上都会有数据, 对于速度的估计会更为精确。

3) 气象方面, 将数据的下载合并到并行后端, 如果数据缺失自动下载。

To Do

- 1) 继续讨论“照片->数据->可视化”的可行性及具体实施方案。
- 2) 并行方面增加磁盘的数据管理和内存的数据管理, 缓存最近使用的数据, 删除长时间未使用的数据。